DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NIJMEGEN The Netherlands

DOUBLE SOBOLEV GRADIENT PRECONDITIONING FOR NONLINEAR ELLIPTIC PROBLEMS

O. Axelsson, J. Karátson

Report No. 0016 (April 2000)

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NIJMEGEN
Toernooiveld
6525 ED Nijmegen
The Netherlands

Double Sobolev gradient preconditioning for nonlinear elliptic problems*

O. Axelsson† J. Karátson‡

Abstract

A mixed variable formulation of a second order nonlinear diffusion problem is shown to lead to a finite element matrix in a product form. This form enables the efficient updating of the nonlinearity in a Picard type iteration method. It is shown that the analysis of convergence of this method is best done in the corresponding Sobolev function space.

1 Introduction

We investigate the numerical solution of the nonlinear diffusion problem

$$\begin{cases} -\text{div } (a(u,\nabla u)\nabla u) = f \\ u_{|\partial\Omega} = 0 \end{cases} \tag{1}$$

on a bounded domain $\Omega \subset \mathbf{R}^N$ with $\inf a > 0$. The function a is assumed to be bounded and piecewise continuous.

The finite element discretization of (??) leads to a nonlinear algebraic system, for whose solution iterative methods are used. The crucial part of the iterative method is most often to find a suitable preconditioner [?]. An important type of preconditioners is based on the Sobolev gradient technique (see [?]), which relies on the underlying Sobolev space properties of the equation and uses discrete Laplacians. The theoretical background for such approaches goes back to [?]. The construction of the theoretical iteration in Sobolev space suggests suitable preconditioners and reveals the convergence properties [?, ?, ?].

In [?] (see also [?]) a mixed variable FEM is proposed for (??), under which the corresponding nonlinear system takes the following form:

$$BM^{-1}(\alpha)B^T\alpha = \beta, \tag{2}$$

where the matrices B, B^T and $M^{-1}(\alpha)$ arise from the discretization of the operators $-\text{div}$, ∇ and the material function $a(u,\nabla u)$, respectively. The product form in

*This research was done during the second author's Hungarian post-doc scholarship Magyary Zoltán.

†Dept. of Mathematics, University of Nijmegen, The Netherlands; axelsson@sci.kun.nl

‡Dept. Applied Analysis, ELTE University, H-1053 Budapest, Hungary; karatson@cs.elte.hu

(??) shows that if one uses Picard type iterations then only the matrix $M^{-1}(\alpha)$ which is symmetric and block diagonal has to be updated. This suggests that the generally faster convergent Newton iteration is not worth the extra work in this case. In addition, except when $a = a(\nabla u)$, the linearized equation that arises in a Newton iteration is nonsymmetric and, therefore, its solution algorithm is more costly. The efficiency of a suitably preconditioned Picard iteration has been experienced indeed in the numerical tests of [?] (see also [?]). This preconditioning will be recalled in the Section 3. (We note that recently a similar preconditioner has been applied to linearized Navier-Stokes equations [?].)

The aim of this paper is to study further the above-mentioned preconditioning. Namely, we wish to carry out analytic investigation of its condition properties. This is based on the corresponding theoretical iteration in Sobolev space, of which the studied sequence is the trace in the FEM subspace. We prove analytic results which reveal the behaviour of this preconditioning.

The paper is organized as follows. Section 2 contains a recapitulation of the derivation of the mixed FEM method and its major properties. In Section 3, the doubly preconditioned method is introduced, and Section 4 contains the estimates on the bounds of the corresponding condition number. Finally in Section 5 we consider the modification of the doubly preconditioned method for problems with several discontinuities.

2 The mixed variable finite element method

Consider the boundary value problem

$$-\nabla \cdot (a\nabla u) = f \ , \ x \in \Omega \subset R^n \tag{3}$$

$$u = g_1 \ , \ x \in \Gamma_D \tag{4}$$

$$a\nabla u \cdot \widehat{n} = g_2 \ , \ x \in \Gamma_N = \Gamma \backslash \Gamma_D \tag{5}$$

where $a = a(u, \nabla u) > 0$ and $\widehat{n}$ denotes the outward pointing normal to Γ_N.

Let $V_N = SPAN(\phi_1, \phi_2, \ldots, \phi_N)$, where ϕ_i are continuous piecewise polynomials of degree k on a finite element mesh on Ω and let $u_h = \sum_{j=1}^{N} \alpha_j \phi_j$, $u_h = g_1$ on Γ_D be the corresponding finite element approximation of u.

We then have

$$\int_\Omega a\nabla u_h \cdot \nabla \phi_i d\Omega = \int_\Omega f\phi_i d\Omega + \int_{\Gamma_N} g_2 \phi_i d\Omega, \ i = 1, 2, \ldots, N$$

or

$$K(u_h \cdot \nabla u_h)\alpha = f^{(1)} + f^{(2)}$$

where $K_{ij} = \int_\Omega a\nabla\phi_j \cdot \nabla\phi_i d\Omega$, $f_i^{(1)} = \int_\Omega f\phi_i d\Omega + \int_{\Gamma_N} g_2\phi_i d\Omega$ and $f^{(2)}$ occurs because of the Dirichlet boundary condition on Γ_D.

In practice we have to use numerical integration on each element and an assembling process for the evaluation of K and $f^{(1)}$. This method suffers from two disadvantages:

(i) If a is a function of u (and of ∇u), updating of K means that the numerical integration and assembling must be repeated.

(ii) Unless lines of discontinuities of a happens to be meshlines, the approximation u_h is less accurate near such lines. Similar considerations are valid in subregions where a has sharp gradients.

The following mixed variable method does not suffer from these disadvantages. Let $z = a\nabla u$. Then (3-5) takes the form

$$\begin{aligned} &\tfrac{1}{a}z - \nabla u = 0 \\ &-\nabla \cdot z = f \\ &u = g_1,\ x \in \Gamma_D \\ &\mathbf{z} \cdot \mathbf{n} = g_2,\ x \in \Gamma_N \end{aligned} \tag{6}$$

For the discretization of (6), we construct finite element subspaces $V \subset H_0^1(\Omega)$ and $W \subset [L^2(\Omega)]^n$. The finite element variational formulation then reads:
Find $u_h \in V_1 \subset H^1(\Omega)$, $u_h = g_1$ on Γ_D and $z_h \in W$ such that

$$\int_\Omega \frac{1}{a} z_h \cdot \tilde{z}_h d\Omega - \int_\Omega \nabla u_h \cdot \tilde{z}_h d\Omega = 0, \forall \tilde{z}_h \in W \tag{7}$$

$$\int_\Omega z_h \cdot \nabla v_h d\Omega = \int_\Omega f v_h d\Omega + \int_{\Gamma_N} g_2 v_h d\Gamma, \forall v_h \in V. \tag{8}$$

By $u_h = g_1$ we mean that $u_h = g_1$ at the nodepoints on Γ_D, but in general an approximation of g_1 on Γ_D, derived for instance by using isoparametric elements.

The variational formulation in (7,8) leads to a nonlinear algebraic system in the form

$$\begin{bmatrix} M^{(1)} & 0 & -B^{(1)^T} \\ 0 & M^{(2)} & -B^{(2)^T} \\ B^{(1)} & B^{(2)} & 0 \end{bmatrix} \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \\ \alpha \end{bmatrix} = \begin{bmatrix} g^{(1)} \\ g^{(2)} \\ f \end{bmatrix}$$

where on each element e_j,

$$\begin{aligned} M_{ij}^{(1)} &= M_{ij}^{(2)} = \textstyle\int_{e_j} a^{-1}\psi_i\psi_j d\Omega \\ B_{ij}^{(1)} &= \textstyle\int_{e_j} \psi_j \frac{\partial}{\partial x}\phi_i d\Omega \\ B_{ij}^{(2)} &= \textstyle\int_{e_j} \psi_j \frac{\partial}{\partial y}\phi_i d\Omega \\ f_i &= \textstyle\int_{\Omega_i} f\phi_i d\Omega + \int_{\Omega_i \cap \Gamma_N} g_2\phi_i d\Gamma \end{aligned}$$

and $g^{(1)}$, $g^{(2)}$ depend on the given Dirichlet dataand ϕ_i, ψ_i are basis functions in V and W, respectively.

The equations may be written in the form

$$\begin{bmatrix} M & -B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \begin{bmatrix} g \\ f \end{bmatrix}$$

and by elimination of β we get

$$BM^{-1}B^T\alpha = f - BM^{-1}g \tag{9}$$

or

$$[B^{(1)}M^{(1)-1}B^{(1)^T} + B^{(2)}M^{(2)-1}B^{(2)^T}]\alpha = f - B^{(1)}M^{(1)-1}g^{(1)} - B^{(2)}M^{(2)-1}g^{(2)}. \tag{10}$$

For the evaluation of $M_{ij}^{(k)}$, $k = 1, 2$, in general we have to use numerical integration. Hereby we can use the rule (exact for quadratic polynomials)

$$\int_{a_1} f d\Omega = \frac{1}{2} area(e_1) \sum_{j=4}^{6} f(N^{(j)}) \tag{11}$$

where $N^{(j)}$, $j = 4, 5, 6$ are the mid-edge points, and then we get

$$M_{ij}^{(k)} = \frac{1}{3} area(e_j) a^{-1}(N^{(j)}) \delta_{ij}, \; k = 1, 2$$

where δ_{ij} is the Kronecker delta function and $N^{(j)}$ is the point associated to ψ_j i.e. the point where $\psi_j = 1$. We observe that $M^{(k)}$, $k = 1, 2$ are diagonal matrices which simplifies the updating and thus the iterative solution method.

In general, for triangles and p-point quadrature rules, where $p = \frac{k(k+1)}{2}$ for a k-degree finite element approximation, $M^{(1)}$ and $M^{(2)}$ (and hence M) become diagonal if Lagrangean basis functions (ψ_i) are associated to the quadrature-points.

As has been shown in [?, ?] this formulation is identical to the primal formulation when we use piecewise linear or piecewise quadratic finite element approximations for u. In general, however, for more accurate numerical integrations, the primal and mixed formulations do not lead to identical approximations for u. For such more accurate approximations, the matrices $M^{(1)}$, $M^{(2)}$ are not diagonal but block diagonal. Hereby, the coefficient 'a' appear as harmonic averages.

For problems with highly varying coefficients, a much coarser mesh can then be used for the mixed variable method for the same order of accuracy as for the primal method.

Numerical experiments in [?] show that, for solving the arising nonlinear system of equations, a simple preconditioned Picard's method is highly competitive as compared to the somewhat more complicated methods of Newton's type. A preconditioning based on spectral equivalence can be used for higher order finite elements in order to simplify the solution of the linear system of equations to be solved in each iteration.

In the present paper estimates of the rate of convergence based on condition numbers are derived to demonstrate this.

3 Construction of the iterative method

3.1 The algorithm using double preconditioning

In [?] an iteration using double preconditioning is proposed for the solution of (??). Namely, the algorithm is as follows: assume that $\alpha^{(m)}$ is obtained, and let $M :=$

$M(\alpha^{(m)})$. Then

$$\alpha^{(m+1)} := \alpha^{(m)} - \tau_m \zeta^{(m)}, \quad \text{where} \tag{12}$$

$$\zeta^{(m)} := (BB^T)^{-1} BMB^T (BB^T)^{-1} (BM^{-1}B^T \alpha^{(m)} - b). \tag{13}$$

Double preconditioning means that $(BB^T)^{-1}$ is applied twice in (??-??).

Using the notation $B^+ := (BB^T)^{-1}B$ for the generalized inverse, (??) is written as

$$\zeta^{(m)} = B^+ M (B^+)^T (BM^{-1}B^T \alpha^{(m)} - b).$$

Here $B^+B^T = I$, $B(B^+)^T = I$.

Remark 1 Before turning to the Sobolev space formulation, we note the following fact. Since

$$\int_\Omega u \cdot \nabla v \;=\; \int_\Omega (-\mathrm{div}\; u) v \qquad (u \in H^1(\Omega)^N, v \in H_0^1(\Omega)),$$

the operator $\nabla : H_0^1(\Omega) \to L^2(\Omega)$ satisfies

$$\nabla^*{}_{|H^1(\Omega)^N} \;=\; -\mathrm{div}\;.$$

This means that the matrices B and B^T inherit the adjoint property from the function space setting.

3.2 The corresponding algorithm in Sobolev space

The analogue of (??-??) can be formulated directly for the strong form of (??). From this we turn to the weak formulation, which is more favourable for both realization and estimates.

We use throughout this paper the inner product

$$\langle u, v \rangle_1 \;:=\; \int_\Omega \nabla u \cdot \nabla v$$

in $H_0^1(\Omega)$, and denote the corresponding norm by $\|u\|_1$.

(a) Strong formulation

Assume that Ω is convex or $\partial\Omega \in C^2$. Then the operator

$$-\Delta : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega) \tag{14}$$

is bijective [?]. In the sequel $-\Delta$ is understood with this domain. Further, if $b \in W^{1,\infty}(\Omega)$ is a given function, then we introduce the notation

$$\hat{L}_{(b)} u := -\mathrm{div}\,(b \nabla u)\,. \tag{15}$$

In the strong formulation of the algorithm we temporarily assume that a is piecewise continuously differentiable.

The algorithm is as follows. Let $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$, assume that $u_m \in H^2(\Omega) \cap H_0^1(\Omega)$ is obtained, and let $a := a(u_m, \nabla u_m)$. Then

$$\begin{aligned} &u_{m+1} := u_m - \tau_m z_m, \quad \text{where} \\ &z_m := (-\Delta)^{-1} \left(-\mathrm{div}\, \frac{1}{a} \nabla \left[(-\Delta)^{-1} \left(-\mathrm{div}\, a \nabla u_m - f \right) \right] \right) \end{aligned} \tag{16}$$

and $\tau_m > 0$.

Using the notation (??), we can write

$$z_m = (-\Delta)^{-1} \hat{L}_{(1/a)} (-\Delta)^{-1} (\hat{L}_{(a)} u_m - f)\,. \tag{17}$$

Owing to (??), the function $z_m \in H^2(\Omega) \cap H_0^1(\Omega)$ is well-defined since

$$\begin{aligned} &w_m := \hat{L}_{(a)} u_m - f \in L^2(\Omega) \;\Rightarrow\; v_m := (-\Delta)^{-1} w_m \in H^2(\Omega) \cap H_0^1(\Omega) \;\Rightarrow \\ &x_m := \hat{L}_{(1/a)} v_m \in L^2(\Omega) \;\Rightarrow\; z_m := (-\Delta)^{-1} x_m \in H^2(\Omega) \cap H_0^1(\Omega). \end{aligned}$$

Using the above notations, z_m is determined by

$$\begin{cases} -\Delta v_m \;=\; \hat{L}_{(a)} u_m - f \\ \quad w_{m|\partial\Omega} \;=\; 0\,, \end{cases} \tag{18}$$

$$\begin{cases} -\Delta z_m \;=\; \hat{L}_{(1/a)} v_m \\ \quad z_{m|\partial\Omega} \;=\; 0\,. \end{cases} \tag{19}$$

(b) Weak formulation

In the weak formulation it suffices to let $u_0 \in H_0^1(\Omega)$ to define the iteration. Assuming again that $u_m \in H_0^1(\Omega)$ is obtained and letting $a := a(u_m, \nabla u_m)$, we define

$$u_{m+1} := u_m - \tau_m z_m, \tag{20}$$

where z_m is determined from the weak formulation of (??)-(??):

$$\int_\Omega \nabla v_m \cdot \nabla h \;=\; \int_\Omega (a\, \nabla u_m \cdot \nabla h - f h) \qquad (h \in H_0^1(\Omega)), \tag{21}$$

$$\int_\Omega \nabla z_m \cdot \nabla h \;=\; \int_\Omega \frac{1}{a}\, \nabla v_m \cdot \nabla h \qquad (h \in H_0^1(\Omega)). \tag{22}$$

4 Conditioning analysis

The aim of this paper is to study the condition properties of the sequence (??-??). That is, introducing the notation

$$C := B^+ M (B^+)^T B M^{-1} B^T, \tag{23}$$

we write (??) as

$$\zeta^{(m)} = C\alpha^{(m)} - \beta, \tag{24}$$

where $\beta = B^+M(B^+)^Tb$, and we investigate $cond(C)$. The experiments in [?] show that the algorithm (??-??) is particularly effective for almost constant coefficients a, i.e. if

$$M \approx c_1 I$$

with some $c_1 > 0$. On the other hand, this allows at the same time discontinuity of a in some points. This is due to the construction of the matrix M as a suitable harmonic average of a. In order to have a heuristic idea as to how this preconditioning improves the condition number of the original matrix $BM^{-1}B^T$, let us first consider the case $M = c_1 I$. Then the following is seen:

(i) The condition number of $BM^{-1}B^T = \frac{1}{c_1}BB^T$ tends to ∞ (proportional to h^{-2}) if the discretization width h tends to 0, since BB^T corresponds to the discrete Laplacian.

(ii) $C = (BB^T)^{-1}BB^T(BB^T)^{-1}BB^T = I$, i.e. $cond(C) = 1$.

In the sequel we consider, without loss of generality, the case $c_1 = 1$, i.e. now $M \approx I$, and on the continuous level $a \approx 1$.

Suggested by the above consideration, our aim is to verify that $M \approx I$ implies $cond(C) \approx 1$. To put it more precisely, if we have a family of problems for some sequence of functions a such that $\|I - M\| \to 0$, then $\|I - C\| \to 0$ and $cond(C) \to 1$.

Our investigations rely on the Sobolev space formulation of the previous section, which shows that (??-??) is the suitable projection of the sequence (??-??) in the corresponding FEM subspace. Hence the conditioning of (??-??) inherits that of the function space case (being asymptotically the same under refinement). Consequently, we first investigate the Sobolev space setting: assuming that $\|a - 1\|_\infty \to 0$, we prove that the condition number of the corresponding operator tends to 1. Based on this, we verify the corresponding result for the matrix C.

4.1 The Sobolev space case

We consider the strong form (??) of the iteration in subsection 1.2 for the formulation of the conditioning property. (This serves a better understanding. Otherwise, the result holds equally for the weak operator, and is in fact proved using the weak formulation.)

Analogous to (??), we define the linear operator $C : H_0^1(\Omega) \supset\to H_0^1(\Omega)$ with domain $D(C) := H^2(\Omega) \cap H_0^1(\Omega)$ by

$$Cu \;=\; (-\Delta)^{-1}\hat{L}_{(1/a)}(-\Delta)^{-1}\hat{L}_{(a)}u\,. \tag{25}$$

That is, in (??) we can write $z_m = Cu_m - \varphi$, where $\varphi = (-\Delta)^{-1}\hat{L}_{(1/a)}(-\Delta)^{-1}f$. We use notation I for the identity operator in $H_0^1(\Omega)$.

Proposition 1 *Let $g := 1 - a$ and assume that*

$$\gamma := \|g\|_\infty < 1.$$

Then

$$\|I - C\| \leq \frac{2\gamma}{1-\gamma}\,. \tag{26}$$

Proof. We have

$$(I-C)u \;=\; (-\Delta)^{-1}\left(-\Delta u - \hat{L}_{(1/a)}(-\Delta)^{-1}\hat{L}_{(a)}u\right)\,,$$

that is,

$$(I-C)u = y$$

where

$$\begin{cases} -\Delta x \;=\; \hat{L}_{(a)}u \\ \quad x_{|\partial\Omega} \;=\; 0\,, \end{cases} \tag{27}$$

$$\begin{cases} -\Delta y \;=\; -\Delta u - \hat{L}_{(1/a)}x \\ \quad y_{|\partial\Omega} \;=\; 0\,. \end{cases} \tag{28}$$

In weak form:

$$\int_\Omega \nabla x\cdot\nabla h \;=\; \int_\Omega a\;\nabla u\cdot\nabla h = \int_\Omega (1-g)\;\nabla u\cdot\nabla h\,, \tag{29}$$

$$\int_\Omega \nabla y\cdot\nabla h \;=\; \int_\Omega \nabla u\cdot\nabla h \;-\; \int_\Omega \frac{1}{a}\;\nabla x\cdot\nabla h = \int_\Omega \nabla u\cdot\nabla h \;-\; \int_\Omega \frac{1}{1-g}\;\nabla x\cdot\nabla h \tag{30}$$

for all $h\in H_0^1(\Omega)$.

Then

$$\|x\|_1 = \|\nabla x\|_{L^2(\Omega)} \leq (1+\gamma)\|u\|_1\,.$$

Further,

$$\int_\Omega \nabla y\cdot\nabla h \;=\; \int_\Omega \nabla u\cdot\nabla h \;-\; \int_\Omega \nabla x\cdot\nabla h \;-\; \int_\Omega \frac{g}{1-g}\;\nabla x\cdot\nabla h$$

and

$$\int_\Omega \nabla x\cdot\nabla h \;=\; \int_\Omega \nabla u\cdot\nabla h \;-\; \int_\Omega g\;\nabla u\cdot\nabla h$$

imply

$$\int_\Omega \nabla y\cdot\nabla h \;=\; \int_\Omega g\;\nabla u\cdot\nabla h \;-\; \int_\Omega \frac{g}{1-g}\;\nabla x\cdot\nabla h\,,$$

hence

$$\|y\|_1 \leq \gamma\|u\|_1 + \frac{\gamma}{1-\gamma}\|x\|_1 \leq \left(\gamma + \frac{\gamma}{1-\gamma}(1+\gamma)\right)\|u\|_1 = \frac{2\gamma}{1-\gamma}\|u\|_1\,.$$

Corollary 1 *(i) If* $\gamma = \|a-1\|_\infty < \frac{1}{3}$, *then* $\|I-C\| < 1$, C *is invertible and*

$$\|C\|\|C^{-1}\| \le \frac{1+\gamma}{1-3\gamma}\,.$$

(ii) If $\gamma \to 0$, *then* $\|I-C\| \to 0$ *and* $\|C\|\|C^{-1}\| \to 1$.

Proof. $\gamma < \frac{1}{3}$ implies $\frac{2\gamma}{1-\gamma} < 1$, hence by (??) $\|I-C\| < 1$ and hence C is invertible. (??) also implies that for any $u \in H_0^1(\Omega)$

$$\frac{1-3\gamma}{1-\gamma}\|u\|_1 = \left(1-\frac{2\gamma}{1-\gamma}\right)\|u\|_1 \le \|Cu\|_1 \le \left(1+\frac{2\gamma}{1-\gamma}\right)\|u\|_1 = \frac{1+\gamma}{1-\gamma}\|u\|_1\,,$$

which proves the estimate for $\|C\|\|C^{-1}\|$. The assertions of (ii) are straightforward.

Remark 2 Assumption $\gamma < \frac{1}{3}$ in Corollary ?? means that $\inf a > \frac{2}{3}$, $\sup a < \frac{4}{3}$. Accordingly, if we consider the case $a \approx c_1$ instead of $a \approx 1$, then $\|I-C\| < 1$ remains valid if $\sup a < 2\inf a$. (In this case the analogues of the above estimates involve $\gamma := \|\frac{a}{c_1} - 1\|_\infty$ instead of $\gamma = \|a-1\|_\infty$.)

The assumption $\sup a < 2\inf a$ can be eliminated or relaxed by changing the Laplacian preconditioner to a suitable piecewise constant variable-coefficient operator. This relies on decomposing the domain Ω in parts Ω_i such that on each Ω_i, $M_i := \sup a_{|\Omega_i} < 2\inf a_{|\Omega_i} =: 2m_i$. The appropriate construction is also to be used if a has several discontinuities (or sharp gradients), and will be discussed in the last section.

Example 1. (a) The following nonlinearity was considered in [?] (problem P2):

$$a(u) = 6 + \frac{10}{\pi} arctan\,\alpha(u_0 - u)\,, \tag{31}$$

where $\alpha > 0$ and $u_0 \in \mathbf{R}$ are constants, and numerical tests were presented for different choices of α and u_0.

Creating a class of functions a by varying α in (??), it can be seen that arbitrarily large gradients may occur. (We may choose $u_0 = 0$.) Namely, let us consider our iteration (??-??), let u_m be fixed and $a = a(u_m)$. Then

$$|\nabla a| = \alpha\frac{10}{\pi}(1+\alpha^2 u_m^2)^{-1}|\nabla u_m|.$$

If $x_0 \in \partial\Omega$, then $u_m(x_0) = 0$ and $|\nabla u_m(x_0)| \neq 0$, hence

$$|\nabla a(x_0)| = \alpha\frac{10}{\pi}|\nabla u_m(x_0)| \to \infty \quad \text{as} \quad \alpha \to \infty\,.$$

On the other hand, Proposition ?? shows that the conditioning of the problem is determined by the deviation of a from its average, and the latter is bounded as $\alpha \to \infty$ owing to the *arctan* function. (The effective estimate of the conditioning in

this case relies on Remark ??. Further, note that with $\alpha \to \infty$ a approaches a sign function.)

(b) Using a slight modification of (??), it can be verified that the condition number may even tend to 1 while the gradient grows arbitrarily high. Namely, we redefine a as

$$a(u) = 6 + \alpha^{-1/2}\frac{10}{\pi} arctan\, \alpha(u_0 - u)\,,$$

and choose again $u_0 = 0$. Letting u_m be fixed, $a = a(u_m)$ and $x_0 \in \partial\Omega$, we obtain as above that

$$|\nabla a(x_0)| = \alpha^{1/2}\frac{10}{\pi}|\nabla u_m(x_0)| \to \infty \quad \text{as} \quad \alpha \to \infty\,.$$

On the other hand, now there holds

$$\|a - c_1\|_\infty = \|a - 6\|_\infty \le 5\alpha^{-1/2} \to 0 \quad \text{as} \quad \alpha \to \infty\,,$$

hence $\|I - C\| \to 0$ (and $\|C\|\|C^{-1}\| \to 1$) for the corresponding doubly preconditioned operator C.

We note that, provided a is piecewise continuously differentiable, the estimate of conditioning for (??) can be given in terms of the integral norm of ∇a instead of $\|a - 1\|_\infty$. (This follows from the representation of the operator $I - C$ below, which shows that the double preconditioning factors out ∇a.).

Proposition 2 *Let $u \in H^2(\Omega) \cap H_0^1(\Omega)$, x as in (??). Then there holds*

$$\langle (I - C)u, v\rangle_1 = \int_\Omega \frac{v}{a^2}\,(a\nabla u - \nabla x)\cdot\nabla a \quad (v \in H_0^1(\Omega)). \tag{32}$$

Proof. Letting $y = (I - C)u$, (??) and (??) yield

$$-\Delta x \;=\; -a\Delta u - \nabla a\cdot\nabla u\,, \qquad -\Delta y \;= -\Delta u + \frac{1}{a}\Delta x + \nabla\frac{1}{a}\cdot\nabla x\,. \tag{33}$$

Hence

$$-\Delta y = -\Delta u + \frac{1}{a}(a\Delta u + \nabla a\cdot\nabla u) + \nabla\frac{1}{a}\cdot\nabla x = \frac{\nabla a}{a}\cdot\nabla u - \frac{\nabla a}{a^2}\cdot\nabla x.$$

Multiplication by v and integration gives (??).

Corollary 2 *Denoting by N the space dimension, let $2 < p < \infty$ if $N = 2$ and let $2 < p \le \frac{2N}{N-2}$ in case $N \ge 3$. Then*

$$\|I - C\| \le 2K_p e^{3\|\log\, a\|_\infty}\|\nabla a\|_{L^q}\,, \tag{34}$$

where $p^{-1} + q^{-1} = 2^{-1}$ and K_p denotes the Sobolev embedding constant (see [?]) in

$$H_0^1(\Omega) \subset L^p(\Omega), \qquad \|u\|_{L^p(\Omega)} \le K_p\|u\|_1 \quad (u \in H_0^1(\Omega)). \tag{35}$$

Proof. (??) implies

$$\langle (I-C)u, v\rangle_1 \le \sup a^{-2}\,(\|a\|_\infty \|u\|_1 + \|x\|_1)\|(\nabla a)v\|_{L^2}\,.$$

Here for any $v \in H_0^1(\Omega)$ with $\|v\|_1 \le 1$, (??) and (??) yield

$$\|x\|_1 \le \|a\|_\infty \|u\|_1 \quad \text{and} \quad \|(\nabla a)v\|_{L^2} \le \|v\|_{L^p}\|\nabla a\|_{L^q} \le K_p \|\nabla a\|_{L^q},$$

whence

$$\|(I-C)u\| \le 2\sup a^{-2}\|a\|_\infty \|u\|_1 K_p \|\nabla a\|_{L^q}\,. \tag{36}$$

Here

$$\sup a \, \sup a^{-2} = e^{\sup \log a - 2\inf \log a} \le e^{3\sup |\log a|},$$

hence (??) implies (??).

Remark 3 As mentioned before, the experiments in [?] exhibit efficiency of the doubly preconditioned iteration if $a \approx 1$ but ∇a may be large in some points. Corollary ?? explains that if $\nabla a \approx 0$ in a large part of Ω such that its big values are concentrated in a set of small Lebesgue measure, then $\|\nabla a\|_{L^q}$ and hence $\|I-C\|$ remains small. In addition, if in a family of problems a is bounded and $\|\nabla a\|_{L^q} \to 0$, then $\|I-C\| \to 0$.

Example 2. Using a second modification of (??), it can be verified that the property described in Remark ?? may hold even if the gradient of a has arbitrarily large pointwise values. The construction (as well as the phenomenon) is analogous to Example 1. Namely, we now redefine a as

$$a(u) = 6 + \alpha^{-q} arctan\,(\alpha^{q+1} u)$$

with some fixed q defined in Corollary ??, and study the behaviour of the gradient as $\alpha \to \infty$. Let again u_m be fixed, $a = a(u_m)$ and $x_0 \in \partial\Omega$. Then

$$|\nabla a(x_0)| = (1+\alpha^{2(q+1)} u_m(x_0)^2)^{-1}\alpha|\nabla u_m(x_0)| = \alpha|\nabla u_m(x_0)| \to \infty \quad \text{as} \quad \alpha \to \infty\,.$$

For simplicity, we now consider $N=1$, $\Omega = (-1,1)$ and assume that u_m is even and $0 \le u_m' \le 1$ on $(-1,0)$. Then

$$\int_\Omega |\nabla a|^q = \int_{-1}^1 |a'|^q = \alpha^q \int_{-1}^1 |u_m'|^q (1+\alpha^{2(q+1)}u_m^2)^{-q} \le \alpha^q \int_{-1}^1 |u_m'|(1+\alpha^{2(q+1)}u_m^2)^{-1} =$$

$$2\alpha^q \int_{-1}^0 u_m' \,(1+\alpha^{2(q+1)}u_m^2)^{-1} = 2\alpha^{-1} arctan(\alpha^{q+1}u_m(0)) \le \pi\alpha^{-1}.$$

Therefore, if $\alpha \to \infty$, then there holds $\|\nabla a\|_{L^q} \to 0$, and consequently $\|I-C\| \to 0$ for the corresponding doubly preconditioned operator C. (Obviously, $\|a-6\|_\infty \to 0$ in this example as well.)

4.2 The discretized case

Let us now consider the matrix $C := B^+M(B^+)^TBM^{-1}B^T$ as in (**??**). As mentioned in the introduction of this section, the matrix C is the suitable projection of the operator (**??**) in the corresponding FEM subspace, hence it inherits the conditioning properties. We formulate and prove this below. The symbol $\langle .,. \rangle$ stands for the Euclidean dot product, further, we use notation

$$\langle x, y\rangle_{B^T} := \langle B^Tx, B^Ty\rangle$$

(also for the corresponding operator norm). Then we have the analogous result (and proof) to that of Proposition **??**:

Proposition 3 *Let $G := I - M^{-1}$. If $\gamma := \|G\| < 1$, then $\|I - C\|_{B^T} \leq \frac{2\gamma}{1-\gamma}$.*

Proof. Using that $I - C = (BB^T)^{-1}\left(BB^T - BMB^T(BB^T)^{-1}BM^{-1}B^T\right)$, we obtain

$$(I - C)z = y,$$

where

$$BB^Tx = BM^{-1}B^Tz, \qquad BB^Ty = BB^Tz - BMB^Tx.$$

The latter are equivalent to

$$\langle B^Tx, B^Th\rangle = \langle M^{-1}B^Tz, B^Th\rangle, \quad \langle B^Ty, B^Th\rangle = \langle B^Tz, B^Th\rangle - \langle MB^Tx, B^Th\rangle$$

for all vectors h. Then the definition of G yields

$$\langle B^Tz, B^Th\rangle = \langle GB^Tz, B^Th\rangle + \langle M^{-1}B^Tz, B^Th\rangle = \langle GB^Tz, B^Th\rangle + \langle B^Tx, B^Th\rangle,$$
$$\langle MB^Tx, B^Th\rangle = \langle G(I-G)^{-1}B^Tx, B^Th\rangle + \langle B^Tx, B^Th\rangle.$$

Subtraction gives

$$\langle B^Ty, B^Th\rangle = \langle GB^Tz, B^Th\rangle - \langle G(I-G)^{-1}B^Tx, B^Th\rangle,$$

whence, using $\|x\|_{B^T} \leq \|M^{-1}\|\|z\|_{B^T} \leq (1+\gamma)\|z\|_{B^T}$, we obtain

$$\|y\|_{B^T} \leq \gamma\|z\|_{B^T} + \frac{\gamma}{1-\gamma}\|x\|_{B^T} \leq \frac{2\gamma}{1-\gamma}\|z\|_{B^T}\,.$$

The assertions of Corollary **??** are also valid for the discrete case, using the B^T norm. (The proof is the same.) Namely:

Corollary 3 *(i) If $\gamma = \|I - M^{-1}\| < \frac{1}{3}$, then $\|I - C\|_{B^T} < 1$, C is invertible and*

$$\|C\|_{B^T}\|C^{-1}\|_{B^T} \leq \frac{1+\gamma}{1-3\gamma}\,.$$

(ii) If $\gamma \to 0$, then $\|I - C\|_{B^T} \to 0$ and $\|C\|_{B^T}\|C^{-1}\|_{B^T} \to 1$.

Remark 4 The assumption $\gamma = \|I - M^{-1}\| < \frac{1}{3}$ can be relaxed or eliminated following Remark **??**. For this $(BB^T)^{-1}$ in (**??**) has to be replaced by $(BWB^T)^{-1}$, where the matrix W is obtained from the piecewise constant weight function w under the discretization. The same construction is to be used if a has several discontinuities. We turn to this in the next section.

5 Modification of the doubly preconditioned method for problems with coefficients with several discontinuities

The construction and the conditioning analysis of the iteration in sections 3-4 rely on the global behaviour of the function a. Clearly, however, this way of estimating the condition number may be too rough. Namely, the estimate using $\gamma = \|a-1\|_\infty$ in Proposition **??** may be deteriorated even by a jump around one point. On the other hand,neither does a large value of γ take into account the possibly simple (e.g. piecewise almost constant) behaviour of a. These considerations imply that our doubly preconditioned method may be improved by involving the local values of a in subdomains of Ω.

To this aim we change the Laplacian preconditioner to a suitable piecewise constant variable-coefficient operator. Namely, the domain Ω is decomposed in parts Ω_i such that on each Ω_i, the oscillation of a is suitably small. Then we define a piecewise constant weight function w such that $w_{|\Omega_i} \equiv c_i$, where c_i is some average of a on Ω_i. The new preconditioner uses formally the operator $\hat{L}_{(w)}u = -\mathrm{div}(w\nabla u)$, namely, (**??**) is replaced formally by

$$Cu \;=\; \hat{L}^{-1}_{(w)}\hat{L}_{(w^2/a)}\hat{L}^{-1}_{(w)}\hat{L}_{(a)}u\,. \tag{37}$$

The weak form of the operator C is derived in the obvious way. This means that the variational problems in the weak form (**??**-**??**) of the algorithm are now posed using the equivalent norm

$$\|u\|_w := \int_\Omega w|\nabla u|^2$$

in $H^1_0(\Omega)$. In the FEM realization this involves certain updating of the stiffness matrices which, however, is not very costly.

Following this modification, the algorithm is defined by

$$u_{m+1} := u_m - \tau_m z_m, \tag{38}$$

where

$$\int_\Omega w\nabla v_m\cdot\nabla h \;=\; \int_\Omega (a\,\nabla u_m\cdot\nabla h - fh) \qquad (h\in H^1_0(\Omega)), \tag{39}$$

$$\int_\Omega w\nabla z_m\cdot\nabla h \;=\; \int_\Omega \frac{w^2}{a}\,\nabla v_m\cdot\nabla h \qquad (h\in H^1_0(\Omega)). \tag{40}$$

(This generalizes (**??**-**??**) in the way described above.)

We use throughout this section the notations

$$M_i := \sup a_{|\Omega_i}, \quad m_i := \inf a_{|\Omega_i} \qquad (i=1,...,k) \tag{41}$$

corresponding to the decomposition of Ω in parts $\Omega_i \quad (i=1,...,k)$.

We formulate and prove below the conditioning properties analogous to Proposition ?? in the following case: the decomposition is chosen such that

$$M_i \;<\; 2m_i \qquad (i = 1, ..., k), \tag{42}$$

and we define w by the arithmetic mean:

$$w_{|\Omega_i} \equiv c_i := \frac{1}{2}(M_i + m_i) \qquad (i = 1, ..., k). \tag{43}$$

Proposition 4 *Let (??) and (??) hold. Then*

(i) $\gamma := \left\| 1 - \frac{a}{w} \right\|_\infty < \frac{1}{3}$;

(ii) $\|I - C\|_w \;\leq\; \frac{2\gamma}{1-\gamma}$;

(iii) $\|I - C\|_w < 1$, *C is invertible and* $\|C\|_w \|C^{-1}\|_w \;\leq\; \frac{1+\gamma}{1-3\gamma}$.

Proof. (i) (??) and (??) yield

$$\frac{3}{4} M_i < c_i < \frac{3}{2} m_i \,. \tag{44}$$

Hence on Ω_i there holds

$$-\frac{1}{3} < 1 - \frac{M_i}{c_i} \leq 1 - \frac{a}{c_i} \leq 1 - \frac{m_i}{c_i} < \frac{1}{3} \,,$$

that is,

$$|1 - \frac{a}{w}| = |1 - \frac{a}{c_i}| < \frac{1}{3} \quad \text{on } \Omega_i \,.$$

(ii) Let $g := 1 - \frac{a}{w}$. There holds

$$(I - C)u = y,$$

where

$$\int_\Omega w \, \nabla x \cdot \nabla h \;=\; \int_\Omega a \, \nabla u \cdot \nabla h = \int_\Omega w(1-g) \, \nabla u \cdot \nabla h \,, \tag{45}$$

$$\int_\Omega w \, \nabla y \cdot \nabla h \;=\; \int_\Omega w \, \nabla u \cdot \nabla h \;-\; \int_\Omega \frac{w^2}{a} \, \nabla x \cdot \nabla h \tag{46}$$

for all $h \in H_0^1(\Omega)$. Hence

$$\|x\|_w \leq (1 + \gamma) \|u\|_w$$

and a similar calculation as in Proposition ?? yields

$$\int_\Omega w \, \nabla y \cdot \nabla h \;=\; \int_\Omega wg \, \nabla u \cdot \nabla h \;-\; \int_\Omega w \frac{g}{1-g} \, \nabla x \cdot \nabla h \,,$$

whence

$$\|y\|_w \leq \gamma \|u\|_w + \frac{\gamma}{1-\gamma} \|x\|_w \leq \frac{2\gamma}{1-\gamma} \|u\|_w \,.$$

(iii) These are the straightforward consequences of (i)-(ii), which follow similarly as in Corollary ??.

Remark 5 Let us now refine the decomposition of Ω such that $M_i < \frac{5}{3} m_i$ holds on each Ω_i instead of (??). Then we may define w by the harmonic average:

$$w_{|\Omega_i} \equiv c_i := \frac{2}{M_i^{-1} + m_i^{-1}} \tag{47}$$

on each Ω_i, such that Proposition ?? remains valid. This is due to preserving the estimate (??).

6 Concluding remarks

It has been shown that nonlinear diffusion problems with rough coefficients can be solved numerically efficiently using a mixed variable variational formulation. Thereby, using a Picard iteration method the costly updating, and assembling of the diffusion matrix required in the primal method is avoided. The preconditioned Picard method works with symmetric and readily updated matrices while a Newton method would require a more expensive method and, generally, a nonsymmetric linearized operator. Still the Picard method can converge arbitrarily fast by proper subdivisions of Ω.

The introduction of the piecewise constant variable-coefficient operator $\hat{L}_{(w)}$ instead of the Laplacian implies that the preconditioner $(BB^T)^{-1}$ in (??) is replaced by a matrix of the form $(BWB^T)^{-1}$, where W is obtained from the discretization of the weight function w. Since w is determined by $a := a(u_m, \nabla u_m)$ in each step of the iteration, this modified preconditioning requires certain updating. However, this is not very costly since the required values corresponding to the same subdomain Ω_i coincide, and suitable element by element or other preconditioning can be used. For further details on solution methods for linear diffusion problems, see [?] and [?].

References

[1] Adams, R.A., *Sobolev Spaces*, Academic Press, 1975.

[2] Axelsson, O., Barker, V.A., *Finite Element Solution of Boundary Value Problems*, Academic Press, 1984.

[3] Axelsson, O., A mixed variable finite element method for the efficient solution of nonlinear diffusion and potential flow equations, in: *Advances in multi-grid methods*, Notes on numerical fluid mechanics, Vol.11 (eds. Braess, D. et al.), pp. 1-11, Braunschweig, 1985.

[4] Axelsson, O., Gustafsson, I., An efficient finite element method for nonlinear diffusion problems, *Bull. Greek Math. Soc.,* 32 (1991), pp. 45-61.

[5] Axelsson, O., *Iterative Solution Methods,* Cambridge University Press, 1994.

[6] Axelsson, O., Chronopoulos, A.T., On nonlinear generalized conjugate gradient methods, *Numer. Math.* 69 (1994), No.1., pp. 1-15.

[7] Elman, H.C., Preconditioning for the steady-state Navier-Stokes equations with low viscosity, *SIAM J. Sci. Comput.,* 20 (1999), No.4, pp. 1299-1316.

[8] FARAGÓ, I., KARÁTSON, J., The gradient–finite element method for elliptic problems, to appear in *Comput. Math. Appls.*

[9] HACKBUSCH, W., *Elliptic Differential Equations (Theory and Numerical Treatment)*, Springer, 1992.

[10] KANTOROVICH, L.V., AKILOV, G.P., *Functional Analysis,* Pergamon Press, 1984.

[11] NEUBERGER, J.W., *Sobolev Gradients and Differential Equations,* Springer, 1997.

[12] RICHARDSON, W.B. JR, Sobolev Gradient Preconditioning for PDE Applications, in: *Iterative Methods in Scientific Computation IV.* (Kincaid, D.R., Elster, A.C., eds.), pp. 223-234, IMACS, New Jersey, 1999.